Original Paper

Human Heredity

Hum Hered 2004;58:18–29
DOI: 10.1159/000081453

Received: May 6, 2004
Accepted after revision: July 27, 2004

A Haplotype-Based Test of Association Using Data from Cohort and Nested Case-Control Epidemiologic Studies

Jinbo Chen[a] Ulrike Peters[b] Charles Foster[c] Nilanjan Chatterjee[a]

[a]Biostatistics and [b]Nutrition Epidemiology Branches, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Md., and [c]Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, Md., USA

Key Words
Nested case-control studies · Haplotype · Genotype · Cox proportional hazards model

Abstract
Haplotype-based risk models can lead to powerful methods for detecting the association of a disease with a genomic region of interest. In population-based studies of unrelated individuals, however, the haplotype status of some subjects may not be discernible without ambiguity from available locus-specific genotype data. A score test for detecting haplotype-based association using genotype data has been developed in the context of generalized linear models for analysis of data from cross-sectional and retrospective studies [1]. In this article, we develop a test for association using genotype data from cohort and nested case-control studies where subjects are prospectively followed until disease incidence or censoring (end of follow-up) occurs. Assuming a proportional hazard model for the haplotype effects, we derive an induced hazard function of the disease given the genotype data, and hence propose a test statistic based on the associated partial likelihood. The proposed test procedure can account for differential follow-up of subjects, can adjust for possibly time-dependent environmental co-factors and can make efficient use of valuable age-at-onset information that is available on cases. We provide an algorithm for computing the test statistic using readily available statistical software. Utilizing simulated data in the context of two genomic regions GPX1 and GPX3, we evaluate the validity of the proposed test for small sample sizes and study its power in the presence and absence of missing genotype data.

Introduction

Population-based association studies are becoming increasingly popular for studying genetic mechanisms of complex diseases. A population-based sample, in which linkage disequilibrium extends over a very short distance of the genome, can permit mapping of disease susceptibility genes on a finer scale than that which can be achieved by family-based linkage studies [2]. Moreover, population-based studies can be powerful tools for identifying susceptibility genes with modest effects that would be difficult to detect in linkage studies [3]. Despite concerns about population stratification, association studies based on unrelated individuals have been appealing because of the cost and other practical advantages. In particular, biologic samples from existing large epidemiologic studies, typically based on unrelated individuals, provide an ex-

KARGER

Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Dr. Jinbo Chen
Biostatistics Branch, Division of Cancer Epidemiology and Genetics
National Cancer Institute, 6120 Executive Blvd EPS 8089
Rockville, MD 20852 (USA)
Tel. +1 301 594 7881, Fax +1 301 402 0081, E-Mail chenjin@mail.nih.gov

cellent opportunity to study susceptibility genes for diseases.

Association studies often involve investigating the risk of a disease in relation to a number of variant single nucleotide polymorphisms (SNPs) identified within a candidate genomic region. Haplotype-based analysis is widely believed to be an efficient way of utilizing such SNP data, as it can fully exploit the multivariate linkage disequilibrium pattern between markers and can potentially capture cis-interactions between causal variants [1, 4, 5]. Statistical considerations also favor haplotype analysis; it can be viewed as a technique for reducing data dimension as the number of haplotypes observed within a genomic region is usually considerably smaller than the number of all possible combination of SNP genotypes [6]. Thus, a haplotype-based model can provide a parsimonious way of specifying disease risk associated with multiple genetic markers and hence can lead to a powerful test for detecting gene-disease association.

In recognition of various advantages of population-based studies and haplotype-based analysis, in recent years, various researchers have developed a number of methods for detecting haplotype-based associations using the case-control epidemiologic study design [1, 4, 7, 8]. This design recruits a fixed number of diseased (cases) and non-diseased subjects (controls) and is efficient for the study of rare diseases. Traditionally, a major concern for the case-control design with questionnaire-based exposure assessment has been the potential bias due to differential recall of exposure history by cases and controls. Moreover, post-disease measurement of endogenous exposures (biomarkers) from biologic samples may not provide accurate information on exposure history before the disease, since biomarkers may be affected by the disease process itself. These considerations, although they do not directly affect genetic association studies, may have important implications for the adjustment of environmental exposures and, more importantly, for studying gene-environment interactions.

An alternative to case-control sampling design that does not have the above-mentioned limitations is the prospective cohort design. In this design, biologic samples and questionnaire-based data are collected at baseline from a group of healthy subjects. This cohort is then followed prospectively for a certain period of time, during which information on the disease incidence, including age at onset, is recorded. Full cohort studies, although popularly used for common traits such as heart disease, are not practical for study of rare diseases such as cancer, as the study may require genotyping and ascertaining expensive environmental exposures for an unnecessarily large number of subjects.

An alternative to prospective cohort design that retains the efficiency advantage of case-control studies is the nested case-control study design [9]. In this design, the biologic sample collected at the beginning of a cohort study is stored for future use. Every time an incident case of the disease occurs in the cohort, a small number (typically one or two) of controls are selected from all subjects who are still under follow-up but have not developed the disease. The genotyping effort and collection of expensive biomarker information are then limited only to cases and the small number of matched controls, thus greatly reducing time, cost, and other practical difficulties associated with full cohort studies. An important aspect of this design is that a control selected at a particular time point remains under follow-up as part of the underlying cohort, and hence remains eligible for being selected as a control for a future case or/and as a future case itself. During analysis, cases and their matched controls are compared with respect to their exposure history using conditional logistic regression method. The possible dependence among matched sets, which may arise due to the fact that the same subject can appear at different matched sets, can be ignored [9].

Existing major cohort studies, such as the Nurses' Health Study [10, 11], European Prospective Investigation into Cancer and Nutrition [12, 13], Multiethnic Cohort Study [14], alpha-tocopherol, beta-carotene Study [15, 16], and Health Professionals Follow-Up Study [17], are now being increasingly used for genetic association studies. Furthermore, the first cohort consortia on breast and prostate cancers combining several thousand cases and controls of existing cohort studies are ongoing to investigate associations between risk of these cancers and variation of about 50 genes in the sex hormone and growth factor pathways. Part of these cohort consortia is the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial initiated by investigators from the US National Cancer Institute to investigate the effectiveness of early detection for these cancers and to identify etiologic determinants of cancer. Besides the cohort consortia, a number of nested case-control studies for various cancers are now being undertaken for evaluating the association of cancer risk with various candidate genes. Motivated from the PLCO study as well as various other ongoing epidemiologic studies, in this article, we develop a haplotype-based method for testing gene-disease association using genotype data from cohort and nested case-control studies.

Traditionally, epidemiologic data from cohort and nested case-control studies are analyzed using survival analysis methods in which incidence and age at onset of the disease are jointly treated as the phenotype of interest, and the 'controls' who do not develop the disease during the follow-up period are treated as being censored. Such methods have several attractive features compared to the alternative logistic regression approach, where disease outcome is treated simply as a binary phenotype describing whether a subject developed the disease or not by the end of the study. In particular, survival analysis methods can account for the fact that 'controls' may be followed for different lengths of time and that controls who are disease-free at a certain age may develop the disease in the future. Moreover, such methods can also efficiently incorporate age at onset information, which can lead to increased power for detecting an association [18].

We assume that the risk (penetrance) of the disease, given a subject's pair of haplotypes (diplotype), is described by the popular Cox proportional hazards model. If one could ascertain which diplotype each subject has, that is, if phase information were known without ambiguity, then testing for association in the Cox model could be performed using standard statistical software. Typically, however, in studies of unrelated subjects with genotype data for individual loci, the haplotype configuration cannot be determined with certainty for all study subjects. Such uncertainty may arise due to intrinsic phase ambiguity associated with genotype data that are heterozygous at multiple loci and/or missing genotype information on some markers. To deal with such phase ambiguity, we propose a score statistic based on the partial likelihood associated with an induced hazard function of the disease given only unphased genotype information. Based on the asymptotic theory, we then show that, under the null hypothesis of no association, the variance of the score function for haplotype effects can be represented in a remarkably simple form. We thus propose a test statistic based on the score and its variance, which can be evaluated using routinely available statistical software. We study the performance of the proposed test using simulated data in the context of two genomic regions, GPX1 and GPX3.

Material and Methods

Cohort and Nested Case-Control Studies

We introduce notations for cohort and nested case-control studies in this section. Let T denote the (potential) age at onset and C denote the (potential) age at censoring. The observed phenotype of an individual consists of the binary disease status $\Delta = I(T \leq C)$ for whether a subject develops disease, where I is the indicator function, and follow-up time $X = \min(T, C)$. In cohort studies, the genotype information G and possibly time-dependent covariates $Z(t)$ are observed for all subjects. Suppose K subjects develop the disease during the course of the study at time points $t_1 < t_2 < \ldots < t_K$. We will define R_k to be the set of all subjects in the cohort who are disease-free just before t_k, including the kth case, and let n_k be the number of subjects in R_k.

For nested case-control studies, $m - 1$ controls are sampled without replacement from non-diseased subjects in R_k, and the study sample consists of all cases in the cohort and their matched controls. Let $\tilde{R}_k$ be the subset of all controls sampled from R_k together with the kth case, which is of size m. The genotype information G and covariates $Z(t)$ are collected only for subjects in $\{\tilde{R}_k: k = 1, \ldots, K\}$. We observe that once the 'risk sets', that is, R_k for cohort studies and $\tilde{R}_k$ for nested case-control studies, have been defined, the standard partial-likelihood analysis (equivalent to the conditional logistic regression) for comparing cases and their matching controls within the 'risk sets' are identical for the two designs. Similar correspondence between cohort and nested case-control designs holds for the methodology we propose below. Thus, for purpose of analysis, the two designs will not be distinguished for the rest of this article. All the methodologies will be described with the risk set for the kth case generally denoted by R_k^*, keeping in mind that $R_k^* = R_k$ for cohort studies and $R_k^* = \tilde{R}_k$ for nested case-control studies.

Haplotype-Specific Hazard Model

We assume throughout this article that the underlying time scale of our analysis is biologic age, but other types of time scale, such as calender year or time since entry into the study, can also be handled without any additional complexity. Let $D = (H_1, H_2)$ denote the diplotype for an individual, that is, the two haplotypes the individual carries in his/her two chromosomes. The hazard function of the disease at age t for an individual in the underlying cohort with diplotype D can be defined as

$$\lambda_D(t) = \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \Pr\{T \in (t, t + \delta t) \,|\, T \geq t, D\}.$$

That is, $\lambda_D(t)$ is the instantaneous probability that an individual with diplotype D will experience the disease at age t given that s/he has been 'at risk', that is, has been free of the disease until time t. Following conventional approach, we specify $\lambda_D(t)$ using the popularly used CPH form:

$$\lambda(t \,|\, D) = \lambda_0(t) e^{\beta^T \psi(D)}, \tag{1}$$

where $\lambda_0(t)$ is the disease hazard associated with a reference diplotype D_0, $\psi(D)$ is a vector of numeric values chosen to represent the diplotype D according to an assumed 'mode of effect', and β is the vector of associated regression coefficients (association parameters). We adopt the convention that if H_0 is the most commonly observed haplotype in the data, $D_0 = (H_0, H_0)$ is then used as the baseline diplotype for the CPH model. The CPH model allows $\lambda_0(t)$, the hazard function associated with the reference diplotype D_0, to be arbitrary (non-parametric) but assumes the hazard ratio $\lambda(t \,|\, D)/\lambda_0(t) = e^{\beta^T \psi(D)}$ to be constant over all ages (t). In association studies, inference on the vector of regression parameters β is of interest, and the baseline hazard function $\lambda_0(t)$ is typically treated as nuisance.

In the above model, the dimension of the vector of regression coefficients as well as their interpretation depend on the choice of the function $\psi(D)$. When testing for the effect of haplotypes is of interest,

$\psi(D)$ can be specified in terms of the constituent haplotypes so that β can represent haplotype-specific hazard ratio (relative risk) parameters. The effect of a pair of haplotypes in a diplotype can be specified according to additive, dominant or recessive models [4, 6, 19]. For example, if there are H observed haplotypes in a dataset and an additive mode of effect is assumed for these haplotypes, the vector $\psi(D)$ would have dimension $(H-1)$ corresponding to $(H-1)$ non-referent haplotypes, each element of which represents the number of copies of the specific non-referent haplotype contained in the diplotype D. In this case, the vector of regression coefficients β is also $(H-1)$-dimensional, and the regression coefficient corresponding to a specific haplotype would represent the relative risk associated with one copy of that haplotype. The main goal of the current article is to develop a score test for simultaneously testing if all of the regression coefficients corresponding to a given mode of effect, $\psi(D)$, are equal to zero, that is, $\beta = 0$. We observe that whatever mode of effect has been chosen, the hypothesis $\beta = 0$ corresponds to a 'global null hypothesis' of no association in the whole genomic region [20]. The statistical power of such test, however, would depend on how well the assumed mode of effect can approximate the true underlying effects of the diplotypes.

If it is desirable to adjust for additional covariates, model (1) can be expanded to

$$\lambda\{t \mid D, Z(t)\} = \lambda_0(t) e^{\beta^T \psi(D) + \gamma^T Z(t)}, \tag{2}$$

where $Z(t)$ denotes a set of possibly time-dependent covariates and γ denotes the vector of associated regression coefficients representing the effect of $Z(t)$ on the disease hazard.

Testing $\beta = 0$ in the Absence of Phase Ambiguity

To facilitate later discussion, we first provide the score statistic when haplotype configurations for all subjects can be ascertained without ambiguity. In this setting, inference on the regression parameters β and γ involved in model (2) can be based on the well-known Cox's partial likelihood function

$$L(\beta, \gamma) = \prod_{k:\Delta_k = 1} \frac{e^{\beta^T \psi(D_k) + \gamma^T Z_k(t_k)}}{\Sigma_{l \in R^*_k} e^{\beta^T \psi(D_l) + \gamma^T Z_l(t_k)}}. \tag{3}$$

In formula (3), the term corresponding to the k-th disease event can be viewed as the probability of the observed disease configuration within the set $\tilde{R}_k$, conditional on the fact that one subject within this set is known to have developed the disease. An appealing feature of the partial likelihood is that it does not depend on the 'nuisance' baseline hazard function $\lambda_0(t)$.

The score function for β based on the partial likelihood (3) at $\beta = 0$ is given by

$$\frac{\partial \log L}{\partial \beta} = U_\beta(\gamma) = \sum_{k:\Delta_k = 1} \left\{ \psi(D_k) - \frac{\Sigma_{l \in R^*_k} \psi(D_l)\, e^{\gamma^T Z_l(t_k)}}{\Sigma_{l \in R^*_k} e^{\gamma^T Z_l(t_k)}} \right\}, \tag{4}$$

based on which a score test for $\beta = 0$ can be constructed. Let $\hat{\gamma}$ denote the maximum partial likelihood estimator of γ at $\beta = 0$ satisfying the score equation $U_\gamma(\gamma) = 0$ where

$$\frac{\partial \log L}{\partial \gamma} = U_\gamma(\gamma) = \sum_{k:\Delta_k = 1} \left\{ Z_k(t_k) - \frac{\Sigma_{l \in R^*_k} Z_l(t_k)\, e^{\gamma^T Z_l(t_k)}}{\Sigma_{l \in R^*_k} e^{\gamma^T Z_l(t_k)}} \right\}. \tag{5}$$

Let $I_{\beta\beta}(\gamma) = \partial \log L / \partial\beta\partial\beta^T$, $I_{\beta\gamma}(\gamma) = \partial \log L / \partial\beta\partial\gamma^T$ and $I_{\gamma\gamma}(\gamma) = \partial \log L / \partial\gamma\partial\gamma^T$, all of which are evaluated at $\beta = 0$. From standard theory of partial likelihood inference, it follows that under the null hypothesis of no association ($\beta = 0$), the test statistic of the form $U_\beta(\hat{\gamma}) S(\hat{\gamma})^{-1} U_\beta(\hat{\gamma})^T$, with

$$S(\hat{\gamma}) = I_{\beta\beta}(\hat{\gamma}) - I_{\beta\gamma}(\hat{\gamma}) I^{-1}_{\gamma\gamma}(\hat{\gamma}) I^T_{\beta\gamma}(\hat{\gamma}), \tag{6}$$

follows a χ^2 distribution with degrees of freedom the same as the dimension of β. When no additional covariate effects (γ) are included in the model, the score function (4) reduces to

$$U_\beta = \sum_{k:\Delta_k = 1} \left\{ \psi(D_k) - \frac{1}{n_k} \sum_{l \in R^*_k} \psi(D_l) \right\},$$

where n_k is the size of R^*_k and test can be performed using $U_\beta I_{\beta\beta}{}^{-1} U^T_\beta$.

The Induced Hazard Model and Partial Likelihood

Although some molecular technologies are now available for determining diplotype (D) status of individuals, technical difficulties and high cost associated with these methods prohibit use of them in large-scale association studies. Instead, in typical epidemiologic studies, the multi-locus genotype information (G) is available. For individuals with multiple heterozygous sites, the phase information, that is, how the alleles are arranged in the two chromosomes, can be ambiguous. Below, we describe how the diplotype-specific risk model (2) can be used to develop a score test for detecting association when only unphased genotype data are available.

We first derive the hazard function for disease induced by the haplotype-specific model (2) given a subject's genotype. Let D_G denote the set of all diplotypes that are consistent with an observed multi-locus genotype G and $\tilde{Z}(t)$ denote the history of $Z(t)$ up to time t. Given model (2), the disease hazard at time t for a subject with genotype G and covariates $Z(t)$ can be expressed in the form (details shown in Appendix A.1)

$$\lambda\{t \mid G, \tilde{Z}(t)\} = \lambda_0(t) r_G\{t; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}, \tag{7}$$

where

$$r_G\{t; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\} = e^{\gamma^T Z(t)} \frac{\Sigma_{D \in D_G} e^{\beta^T \psi(D)} \exp\left[-\int_0^t \lambda\{s \mid D, \tilde{Z}(s)\} ds\right] \mathrm{pr}_{\mathbf{f}}(D)}{\Sigma_{D \in D_G} \exp\left[-\int_0^t \lambda\{s \mid D, \tilde{Z}(s)\} ds\right] \mathrm{pr}_{\mathbf{f}}(D)}$$

and $\mathrm{pr}_{\mathbf{f}}(D)$ denotes the population frequency of the diplotype D computed under a set of haplotype frequencies $\mathbf{f}$ assuming Hardy Weinberg Equilibrium (HWE). Above, $\lambda_0(\cdot)$ represents the baseline hazard as a functional form. We observe that in equation (7), $r_G\{t; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}$ can be viewed as a relative hazard function associated with genotype G and covariates $Z(t)$ in reference to the baseline hazard $\lambda_0(t)$. Unlike the original CPH model for diplotype-specific hazard (equation 2), the induced model for genotype-specific risk does not follow the proportional hazards form, as the relative hazard function $r_G\{t; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}$ depends on t through the function $\lambda_0(t)$.

Based on formula (7) for the induced hazard model, we write down a partial likelihood of the data as

$$L\{\beta, \gamma, \mathbf{f}, \lambda_0(t)\} = \prod_{k:\Delta_k = 1} \frac{r_{G_k}\{t_k; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}}{\Sigma_{l \in R^*_k} r_{G_l}\{t_k; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}}. \tag{8}$$

We observe that unlike the partial likelihood of the data under the standard CPH model (formula 3), L involves not only the association parameters of interest β and covariate effects γ, but also the baseline hazard function $\lambda_0(t)$ and the haplotype frequency parameters $\mathbf{f}$. The score function of β evaluated at $\beta = 0$, however, is free of the nuisance hazard function $\lambda_0(t)$. Below, we propose a test for $\beta = 0$ based on this score function with γ and $\mathbf{f}$ estimated from the data.

The Score Statistic and Its Theoretical Properties

The score function for β corresponding to the partial likelihood (8) under the null hypothesis of $\beta = 0$ is

$$U_\beta(\mathbf{f}, \gamma) = \sum_{k:\Delta_k=1} \left(\mathrm{E}_{\mathbf{f}}\{\psi(D)|G_k\} - \frac{\Sigma_{l\in R^*_k} e^{\gamma^T Z_l(t_k)} \mathrm{E}_{\mathbf{f}}\{\psi(D)|G_l\}}{\Sigma_{l\in R^*_k} e^{\gamma^T Z_l(t_k)}} \right), \quad (9)$$

where $\mathrm{E}_{\mathbf{f}}\{\psi(D)|G\}$, the expected value for the function $\psi(D)$ given a subject's genotype G, is determined according to $\mathrm{pr}_{\mathbf{f}}(D|G) = \mathrm{pr}_{\mathbf{f}}(D)/\Sigma_{D\in D_G} \mathrm{pr}_{\mathbf{f}}(D)$, the probability distribution for all of the subject's possible diplotypes given that the subject has genotype G. Two features of the score function require attention. First, it takes a form similar to the standard partial-likelihood score function shown in formula (4), except that the diplotype function $\psi(D)$ is replaced by its expected value given the genotype data G. Second, under the null hypothesis $\beta = 0$, the score function does not involve the baseline hazard $\lambda_0(t)$, although it still involves the additional parameters $\mathbf{f}$ and γ. To construct the score statistic from $U_\beta(\mathbf{f}, \gamma)$, one thus needs to estimate only $\mathbf{f}$ and γ.

Under the null hypothesis, γ can be estimated by maximizing the partial likelihood (8) over γ with β fixed at zero. The corresponding score function for γ has the standard partial likelihood-score form given in formula (5). Moreover, under the null hypothesis of no genetic association and assuming the independence of genetic susceptibility and other cofactors $Z(t)$, the maximum-likelihood estimate of the haplotype frequencies $\mathbf{f}$, assuming HWE, can be obtained using the EM algorithm [21–23] based on the pooled sample of all subjects in the study. The formula for the corresponding likelihood is given in equation (4) of Excoffier and Slatkin [21]. For studies involving different ethnic groups who may have different genetic backgrounds, haplotype frequencies should be estimated stratified by ethnicity. For nested case-control studies, it is important to note that although an individual subject can appear multiple times in different matched case-control sets, he/she should be entered only once in the EM algorithm for haplotype frequency estimation. We denote estimates of γ and $\mathbf{f}$ under $\beta = 0$ by $\hat{\gamma}$ and $\hat{\mathbf{f}}$, respectively.

Development of the score test based on the function $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ requires a study of its asymptotic properties. A sketch of the main results is presented below. The technical proof of these results relies on martingale theory and is deferred for the appendix A3. Define as before $I_{\beta\gamma}(\mathbf{f}, \gamma)$ and $I_{\gamma\gamma}(\gamma)$ to be $\partial U_\beta(\mathbf{f}, \gamma)/\partial\gamma$ and $\partial U_\gamma(\gamma)/\partial\gamma$, respectively, evaluated at $\beta = 0$. We first show that $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ can be asymptotically represented as

$$\frac{1}{\sqrt{n}} U_\beta(\hat{f}, \hat{\gamma}) \approx \frac{1}{\sqrt{n}} U_\beta(\mathbf{f}, \gamma) - I_{\beta\gamma}(\mathbf{f}, \gamma) I^{-1}_{\gamma\gamma}(\gamma) \frac{1}{\sqrt{n}} U_\gamma(\gamma). \quad (10)$$

In equation (10), the first term corresponds to the score-function for β that could be used if $\mathbf{f}$ and γ were known, and the second term corresponds to an adjustment term that accounts for additional variability due to estimating parameter γ from the data. We observe that estimation of haplotype frequencies $\mathbf{f}$ does not add any additional variability, a consequence of the fact that $n^{-1}\,\partial U_\beta(\mathbf{f}, \gamma)/\partial\mathbf{f}$ is asymptotically negligible (see appendix A3.1 for details). A similar phenomenon has been observed in the generalized linear model framework [1, 8].

Based on the representation (10), we then show that under the null hypothesis of $\beta = 0$, the score-function $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ is asymptotically normally distributed with a variance-covariance matrix of the form $v_{\beta\beta} - i_{\beta\gamma} i^{-1}_{\gamma\gamma} i^T_{\beta\gamma}$ where $v_{\beta\beta}$ is the asymptotic variance of $n^{-1/2}\, U_\beta(\mathbf{f}, \gamma)$, and $i_{\beta\gamma}$ and $i_{\gamma\gamma}$ are the limiting versions of $I_{\beta\gamma}$ and $I_{\gamma\gamma}$, respectively. Clearly, $i_{\beta\gamma}$ and $i_{\gamma\gamma}$ can be estimated from the data by the information matrices $I_{\beta\gamma}$ and $I_{\gamma\gamma}$, respectively. Moreover, we show that under the null hypothesis of no association ($\beta = 0$), $v_{\beta\beta}$ can be estimated by an information matrix $I_{\beta\beta}(\hat{\mathbf{f}}, \hat{\gamma})$ of the same form as $I_{\beta\beta}(\hat{\gamma})$ with $\psi(D)$ being replaced by $\mathrm{E}_{\mathbf{f}}\{\psi(D)|G\}$.

Thus, under the null hypothesis of no genetic association, the test statistic of the form $U_\beta(\hat{\mathbf{f}}, \hat{\gamma}) S^{-1}(\hat{\mathbf{f}}, \hat{\gamma}) U^T_\beta(\hat{\mathbf{f}}, \hat{\gamma})$, where

$$S(\hat{\mathbf{f}}, \hat{\gamma}) = I_{\beta\beta}(\hat{\mathbf{f}}, \hat{\gamma}) - I_{\beta\gamma}(\hat{\mathbf{f}}, \hat{\gamma}) I^{-1}_{\gamma\gamma}(\hat{\gamma}) I^T_{\beta\gamma}(\hat{\mathbf{f}}, \hat{\gamma}), \quad (11)$$

is asymptotically distributed as χ^2 with degrees of freedom the same as the dimension of β. When no covariates are involved, $U_\beta(\mathbf{f}, \gamma)$ reduces to

$$U_\beta(\mathbf{f}) = \sum_{k_k:\Delta_k=1} \left[\mathrm{E}_{\mathbf{f}}\{\psi(D)|G_k\} - \frac{1}{n_k} \sum_{l\in R^*_k} \mathrm{E}_{\mathbf{f}}\{\psi(D)|G_l\} \right],$$

and the test statistic takes the simple form $U_\beta(\hat{\mathbf{f}}) I^{-1}_{\beta\beta}(\hat{\mathbf{f}}) U^T_\beta(\hat{\mathbf{f}})$.

Computation of the Score Statistic

The proposed test statistic can be computed using standard statistical software. By comparing formulas for $I_{\beta\beta}(\gamma)$ and $I_{\beta\beta}(\mathbf{f}, \gamma)$ and comparing formulas for $S(\gamma)$ and $S(\mathbf{f}, \gamma)$, we observe that $I_{\beta\beta}(\mathbf{f}, \gamma)$ and $S(\mathbf{f}, \gamma)$, which correspond to *unknown* phase information, can be obtained from the formulas for $I_{\beta\beta}(\gamma)$ and $S(\gamma)$ corresponding to *known* phase information by simply replacing $\psi(D)$ by $\mathrm{E}_{\hat{\mathbf{f}}}\{\psi(D)|G\}$ throughout. As indicated before, the score function $U_\beta(\mathbf{f}, \gamma)$ also has the same form as $U_\beta(\gamma)$ except that $\psi(D)$ is replaced by $\mathrm{E}_{\hat{\mathbf{f}}}\{\psi(D)|G\}$. Thus, once $\mathrm{E}_{\hat{\mathbf{f}}}\{\psi(D)|G\}$ has been obtained based on the estimates of the haplotype frequencies $\mathbf{f}$, the rest can be handled by any standard statistical software that can perform CPH analysis. Next, we describe how we implemented this in the statistical software Splus/R, which we use for all our numerical examples.

1. Obtain an estimate of the haplotype frequencies $\hat{\mathbf{f}}$ via the EM algorithm [21–23], using, for example, the HAP.EM function in software Splus/R, applied to the pooled sample of all non-duplicated subjects in the study.
2. Based on $\hat{\mathbf{f}}$, compute $\mathrm{E}_{\hat{\mathbf{f}}}\{\psi(D)|G\}$ for each individual subject and assemble them to form a design matrix for genotypes.
3. Perform ordinary Cox analysis using function COXPH, including only $Z(t)$ as covariates to obtain $\hat{\gamma}$.
4. Perform ordinary Cox analysis with $[\mathrm{E}_{\hat{\mathbf{f}}}\{\psi(D)|G\}, Z]$ as covariates, but setting the starting parameter values *INIT* = to be $\beta = 0$ and $\gamma = \hat{\gamma}$ and the maximum number of iterations *ITER.MAX* to be 0.
5. The output score statistic is what is desired if no covariates are involved. Otherwise, the element at the leftmost corner of the output Hessian matrix would be $S^{-1}(\hat{\mathbf{f}}, \hat{\gamma})$. $U_\beta(\hat{f}, \hat{\gamma})$ can be obtained by summing up the individual score function values, which can be obtained using the function COXPH.DETAIL, across observed disease times.

This algorithm is also useful for simplifying the computation for the score test for generalized linear models [1]. In particular, Schaid et al. [1] provided a formula for estimating $v_{\beta\beta}$ in the context of GLM based on the theory of the EM algorithm. However, we observed that the first term in their formula asymptotically converges to zero, so that, once the quantity $\mathrm{E}_p(X_{gi})$, the conditional expectation of the diplotype-specific design matrix given the genotype data, was estimated for each individual, the score statistic could be computed using the standard program for fitting GLM following the same steps described above.

Table 1. Estimated haplotype frequencies for GPX1 and GPX3

Haplotypes	GPX1						GPX3					
	S_1	S_2	S_3	S_4	S_5	freq[b]	S_1	S_2	S_3	S_4	S_5	freq
H_1	0[a]	0	0	0	1	0.298	0	0	0	0	0	0.433
H_2	0	0	0	0	0	0.267	0	0	0	1	0	0.172
H_3	0	0	1	0	1	0.152	0	0	1	0	0	0.116
H_4	1	1	0	1	0	0.117	1	1	1	1	1	0.080
H_5	0	0	1	0	0	0.099	0	1	1	0	0	0.078
H_6	1	0	0	1	0	0.034	0	0	1	1	1	0.045
H_7	1	1	1	0	0	0.032	1	1	1	0	0	0.039
H_8							1	1	0	0	0	0.038

[a] 0/1 are common/rare alleles respectively.
[b] Haplotype frequencies.

Performance Studies

We evaluated the performance of the proposed test using simulated genotype data in the context of two genomic regions, GPX1 and GPX3. We are currently planning a nested case-control study based on the PLCO cohort to investigate the association between risk of prostate cancer and genomic regions encoding four glutathione peroxidase selenoproteins. A pilot project for re-sequencing four genes in this region (GPX1, GPX2, GPX3 and GPX4) using a sample of 31 Caucasian-American subjects which is a part of the SNP500 database has recently been completed. We plan to use this data for selecting tagging SNPs to be genotyped in the main case-control study. For our simulation study, we selected 5 common SNPs (minor allele frequency > 5%) for each of the two genes, GPX1 and GPX3, for the purpose of illustration. Table 1 shows the corresponding haplotype structures and frequencies that we estimated using the re-sequencing data.

We simulated data in a setting where a single SNP in a gene was disease-causative but assumed that the causal SNP was not selected for genotyping in the main association study, a scenario where haplotype analysis based on marker SNPs is often believed to be a powerful approach for detecting association [20]. Based on the haplotype frequencies shown in table 1, we first generated the diplotype data under the HWE assumption for a cohort of subjects. We then generated the disease end point for each subject in the cohort based on a penetrance model, assuming one of the five SNPs was disease-causative and the other four were 'pure markers', in the sense that they did not affect the disease risk given the true causal SNP. In particular, when the ith locus was selected as disease-causative, the time-to-disease-onset (T) for a subject was generated from the hazard model $\lambda(t|S_i) = \lambda_0 e^{\beta A_{S_i}}$, where λ_0 denotes the baseline hazard of the disease (assumed to be constant over time), A_{S_i} denotes the number of variant alleles (A) at the ith locus and β denote the associated hazard ratio parameter that quantifies the increase in hazard of the disease associated with one copy of A. For each subject, we then generated a random censoring time (C) based on an exponential distribution. The disease status was then defined as $\Delta = I(T < C)$, and the follow-up time X was equal to T if the subject was a case ($\Delta = 1$) and equal to C if the subject was a control ($\Delta = 0$). The baseline hazard for the disease (λ_0) and for the censoring time were chosen in such a way that the proportion of cases was fixed approximately at 10% as a fraction of the total cohort size.

Once the data for the full cohort study were generated, we then sampled a nested case-control study within the cohort by selecting all the cases and a set of matched controls. We used a 1:1 case-control matching ratio. For each case, we selected a matched control by randomly sampling a subject from the non-diseased people in the 'risk-set' defined by all subjects in the cohort whose follow-up time X was greater than or equal to that of the case. For analysis of each set of simulated data, we assumed genotype data were available only for the marker SNPs, but not for the causal SNP. To investigate the possible effect of missing genotypes, we further deleted the genotype information for each individual marker for a randomly selected subset of subjects. For computing the score statistic, we fitted the additive model for the effect of haplotypes, that is, we assumed that the relative risk associated with two copies of a haplotype was equal to the square of that associated with one copy. All simulations were repeated for 500 times.

Results

An important theoretical result whose proof appeared in the appendix was that under the null hypothesis of no association, the asymptotic variance of the score statistic $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ was not affected by variability associated with the estimation of the haplotype frequency parameters $\mathbf{f}$ by $\hat{\mathbf{f}}$. In particular, we showed that when no covariates were involved, the asymptotic variance of the score statistic $U_\beta(\hat{\mathbf{f}})$ could simply be estimated by an information matrix $I_{\beta\beta}(\hat{\mathbf{f}})$, given in formula (14) of the appendix (evaluated at $\gamma = 0$). We assessed the small sample performance of this estimator using our simulated data. For this part, we used only data for the GPX1 gene, assuming that the SNP S_1 was disease-causative and S_2, S_3, S_4 and S_5 were markers,

Table 2. Empirical/estimated standard error for haplotype-specific score functions

p[a]	n_{case}[b]	H_1[c]	H_2	H_3	H_4	H_5^*
0	100	0.246/0.254	0.510/0.529	0.397/0.397	0.232/0.240	0.513/0.544
	500	0.109/0.112	0.228/0.228	0.166/0.172	0.101/0.106	0.236/0.236
0.05	100	0.265/0.266	0.532/0.560	0.403/0.420	0.244/0.246	0.523/0.568
	500	0.118/0.117	0.242/0.240	0.182/0.181	0.108/0.108	0.240/0.244
0.10	100	0.273/0.278	0.570/0.593	0.446/0.444	0.243/0.252	0.581/0.598
	500	0.120/0.122	0.253/0.254	0.187/0.192	0.109/0.111	0.247/0.252
0.15	100	0.288/0.291	0.576/0.630	0.462/0.469	0.256/0.258	0.603/0.623
	500	0.128/0.128	0.258/0.267	0.198/0.204	0.113/0.114	0.261/0.264
0.20	100	0.294/0.306	0.627/0.677	0.483/0.504	0.252/0.267	0.594/0.657
	500	0.138/0.135	0.294/0.285	0.219/0.217	0.118/0.117	0.274/0.276

[a] Proportion of missing genotype per SNP.
[b] n_{case} is approximated by 10% of cohort size.
[c] Haplotypes corresponding to those in table 1. H_5^* is the category combining H_5, H_6, and H_7.

which, in turn, defined seven different haplotypes (see table 1). We chose the most common haplotype as the baseline and combined two haplotypes that had frequencies lower than 5% into a single 'rare haplotype' category.

Table 2 lists the results of comparing the empirical and the estimated variances (averaged over simulated data) for the score function corresponding to each individual haplotype. Overall, the asymptotic estimator appeared to perform well. Even when the sample size was small, involving only about 100 cases, the extent of bias for the asymptotic variance estimator was very small. The bias was larger for rarer than for common haplotypes, and it increased with the proportion of missing genotypes. As the sample size increased, all of the biases became negligible.

Next, we examined the performance of the global score test by evaluating its nominal type I error rate and power. It has often been a matter of debate whether and in what situation haplotype-based tests of association could be more powerful than SNP-based tests of association. Thus, for comparison purpose, we also considered a SNP-based global test of association as follows. We used a standard score test procedure for the Cox proportional hazards model to test for the association of the disease with individual marker SNPs, adjusting for multiple comparisons using the Bonferonni procedure. If at least one marker SNP was found to be significantly associated with the disease, then the global null hypothesis of no association in the genomic region was rejected. In the SNP-based test, if a subject had missing genotype for a particular SNP, we imputed the missing genotype by its expected allele count $2\hat{p}^2 + \hat{p}(1 - \hat{p})$, where $\hat{p}$ is the estimated allele frequency from the data using the pooled sample of cases and controls.

A practical issue for haplotype-based association analysis is how to best deal with rare haplotypes so that their involvement in a test procedure does not cause numerical instability and/or loss of power due to the use of large degrees of freedom. We adopted the common practice of combining all haplotypes with estimated frequency lower than 5% into a single 'rare haplotype category'. If the frequency of this combined category was still less than 5%, then we further combined it with the least common haplotype that had an estimated frequency higher than 5%. Since estimates of haplotype frequencies varied across simulated data and rare haplotypes may appear in one simulation but not in another, the test for association for the same genomic region could have different degrees of freedom for different replications. This reflected what would happen in a real study setting if studies were repeated.

To evaluate nominal type I error rates, we used the same simulation setup as that for table 2. From results shown in table 3, we observed that, in most situations both the SNP-based and haplotype-based tests maintained the chosen α level. When the sample size was small, occasionally, the observed type I error rates for both of the tests exceeded the chosen α level.

We evaluated the power of the proposed score test and the global SNP-based test in a variety of different scenarios. We simulated data for both genes GPX1 and GPX3. For each gene, we selected one SNP, in turn, to be disease-

Table 3. Type I error rate for SNP/haplotype-based global test for association

	n_{case}[a]	Overall proportion of missing SNPs				
		0	0.05	0.10	0.15	0.20
0.01	100	0.012[b]/0.007[c]	0.008/0.010	0.008/0.011	0.011/0.006	0.006/0.003
	500	0.008/0.007	0.011/0.006	0.008/0.012	0.009/0.007	0.018/0.014
0.05	100	0.043/0.041	0.058/0.049	0.046/0.049	0.037/0.040	0.048/0.050
	500	0.035/0.047	0.044/0.043	0.040/0.043	0.044/0.040	0.032/0.049
0.10	100	0.107/0.113	0.104/0.100	0.098/0.103	0.112/0.124	0.094/0.099
	500	0.086/0.121	0.089/0.109	0.105/0.119	0.102/0.125	0.102/0.109

[a] n_{case} is approximated by 10% of cohort size.
[b] Empirical type I error rate for the SNP-based analysis.
[c] Empirical type I error rate for the haplotype-based analysis.

Table 4. Power for haplotype-based vs. SNP-based test for global association

Causal SNP		Overall proportion of missing SNPs when causal SNP is not genotyped				
		0	0.05	0.10	0.15	0.20
S1	GPX1	0.906[a]/0.956[b]	0.900/0.908	0.878/0.868	0.872/0.794	0.872/0.746
	GPX3	0.794/0.892	0.792/0.868	0.780/0.772	0.766/0.744	0.730/0.704
S2	GPX1	0.704/0.822	0.698/0.774	0.686/0.714	0.690/0.640	0.654/0.594
	GPX3	0.848/0.918	0.838/0.856	0.806/0.798	0.802/0.732	0.764/0.694
S3	GPX1	0.148/0.174	0.146/0.138	0.148/0.110	0.118/0.082	0.140/0.098
	GPX3	0.814/0.730	0.792/0.666	0.764/0.596	0.746/0.552	0.708/0.476
S4	GPX1	0.748/0.864	0.744/0.810	0.720/0.754	0.722/0.664	0.708/0.618
	GPX3	0.620/0.614	0.592/0.442	0.556/0.334	0.528/0.264	0.484/0.222
S5	GPX1	0.666/0.674	0.670/0.582	0.676/0.504	0.656/0.410	0.656/0.344
	GPX3	0.812/0.802	0.806/0.758	0.766/0.696	0.760/0.664	0.736/0.628

[a] Power for the haplotype-based analysis.
[b] Power for the SNP-based analysis.

causative and used the other four SNPs as markers in both of the haplotype- and SNP-based tests of association. In each scenario, we simulated data assuming $\beta = \log(3)$ so that the disease hazard increased 3-fold for carrying one copy of the true causal SNP. We used cohort size of 1,000 and kept the significance level at 0.05.

Results are shown in table 4. We observed that with 5 SNPs in the two genomic regions within GPX1 and GPX3, the SNP-based test of association had generally similar or more power compared to the haplotype-based test when there were no missing genotype data. Intuitively, these results are somewhat unexpected. However, careful inspection of the linkage-disequilibrium pattern shown in tables 2 and 3 reveals that the causal SNP was tightly linked with at least one of the marker SNPs in the same gene in most of the simulation scenarios. Thus, the test of association using individual SNPs, even when the true causal SNP was not genotyped, had high power.

When we allowed for missing genotype data, we observed that the power for the haplotype-based test increased relative to the SNP-based test for all simulation scenarios. In particular, even with a modest proportion of missing genotype information (10% per marker), there was a significant gain in power for using the haplotype-based approach when the true causal SNP was S_3 or S_5 for GPX1 and S_4 for GPX3. These results suggest that in the presence of significant missing genotype data, haplotype-based tests of association may be optimal even for regions with very high linkage where one would ordinarily expect a SNP-based test of association to be more powerful.

Discussion

In summary, we have developed a haplotype-based score test for detecting the association of a disease with a genomic region of interest using prospective follow-up information and unphased genotype data collected from cohort and nested case-control studies. The proposed method extends the methods of testing association using generalized linear models [1, 8] to the censored-survival-data setting. It can account for differential follow-up for subjects, make efficient use of age at onset information, adjust for the effect of possibly time-dependent covariates and account for individual matching in the nested case-control design. Similar to Schaid et al. [1] and Zaykin et al. [8], the proposed method could be used to test for an association of the disease with the whole genomic region as well as with individual haplotypes. The computational algorithm we provide takes advantage of existing software and thus requires very limited extra computing effort.

We have developed the test procedure assuming HWE; this assumption, however, is not essential. An examination of the formula for $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ (equation 9) shows that the HWE assumption is utilized only to compute $E_{\mathbf{f}}\{\psi(D)|G\}$, the conditional expectation of the diplotype-specific design vector given the genotype data. We observe that the score statistic remains unbiased under the null hypothesis of no association even if we replace $E_{\mathbf{f}}\{\psi(D)|G\}$ by $\varphi_{\mathbf{f}}(G)$, any arbitrary function of the genotype data G. Moreover, from our theoretical calculations it can also be seen that the asymptotic variance of such a modified score statistic could be derived using the same formulas as those derived assuming HWE by replacing $E_{\mathbf{f}}\{\psi(D)|G\}$ by $\varphi_{\mathbf{f}}(G)$ throughout. Thus, the proposed score test remains a valid test even when HWE assumption is violated: $E_{\mathbf{f}}\{\psi(D)|G\}$ computed under HWE can be viewed as a particular function for G that is not necessarily the true conditional expectation.

We assumed no ties between the ages at onset of the cases. For standard Cox analysis of cohort data, both exact and approximate solutions are available for dealing with the ties in t_k [24, 25]. Since all of the quantities involved in the proposed test statistic can be obtained based on a standard CPH analysis, ties could be handled in this procedure by applying the standard solutions. The formal justification is straightforward and is not provided here. For nested case-control studies, following Borgan et al. [26], we suggest randomly breaking ties before the analysis is performed. We also observe that the proposed test procedure, although it has been derived in the context of a CPH model for prospective designs, is also suitable for conditional logistic regression analysis of matched case-control studies.

Using our proposed test, we evaluated the effect of missing marker genotype data on the power of global tests in the context of nested case-control studies. We found that missing genotype data led to a much greater loss of power for the SNP-based test than for the haplotype-based test. This result suggests that haplotypes, which exploit the multivariate correlation structure among the available markers, can efficiently recover missing genotype information on individual markers. Of course, in the SNP-based test, one can also recover missing genotype information on a specific SNP by utilizing data available on other SNPs. In particular, the missing genotype data for a specific SNP can be imputed by the conditional expectation of the number of variant alleles given the genotype for the other SNPs, a quantity that can be computed based on the estimated haplotype frequencies under the HWE assumption. In addition, an alternative approach for evaluating the association of multiple SNPs with a disease could be to test the significance of coefficients in a multivariate genotype-based regression model that does not require phase information [33]. The power of such a test procedure, however, could be low compared to a haplotype analysis if the association primarily exists due to cis-interaction between SNPs.

A naive, but simple, approach for analyzing cohort-based studies could be performing a standard logistic regression analysis, where each subject is classified as a 'case' or a 'control' depending on whether the subject developed the disease by the end of follow-up period or not. In certain situations, for example, when the underlying baseline disease hazard is constant over time and the censoring is completely random, the null value for relative risk parameters in the Cox model may correspond to the null value for odds ratio parameters in the induced logistic regression model. Consequently, in these situations, testing for haplotype effects using the logistic regression method should be valid in the sense that it would have correct α level. To check the above assertion, we conducted a small scale simulation study in the scenario of table 3 and found that the naive logistic regression analysis indeed maintained the α level in this situation (data not shown). However, when we compared power in the scenario of table 4, we found in several situations there was substantial loss of power in the logistic regression analysis compared to the Cox analysis (data not shown).

Lin [27] recently proposed methods for testing and estimation of haplotype effects using data from cohort studies. They proposed modelling the effect of haplotypes

on disease risk using a proportional hazards model similar to ours. For inferences in the presence of phase ambiguity, however, the author considered a likelihood-based approach. As described above, the score test we proposed based on the induced hazard function of the disease given the genotype data, has several practical advantages in terms of general applicability to alternative designs and computational simplicity. The likelihood-based approach of Lin, on the other hand, allows both testing and parameter estimation in the setting of the full cohort design. In a companion paper in preparation, we are developing a method for estimating the haplotype-specific risk parameters (β) in the Cox model. We have found that the induced hazard function and the corresponding partial likelihood formula that we have utilized in this article for developing a test procedure are also useful for developing relatively simple estimation methods for various alternative study designs.

When evaluating the genetic determinants of a disease, some preliminary knowledge may exist about possibly time-dependent effects of susceptibility genes. Familial aggregation studies, for example, suggest that heritable factors increase the risk of breast cancer more strongly at younger ages than at older ages. In such situations, more powerful test procedures can be obtained by weighting the contribution of each risk set to the score statistic (equation 9) according to a time-dependent weight function. In particular, a test for the age-stratified effect of a gene may be constructed based on simple weight functions of the form $w(T) = I(a < T \leq b]$ where $(a, b]$ is a fixed age-range of interest. Weighted forms of the score statistic can also be applied to reduce the influence of outlying observations. In the context of standard survival analysis, various approaches for weighting the log-rank statistics [24, 28] are available. The utility of these weighting methods in the context of genetic association studies requires further research.

Other future areas of research include extension of the proposed method to the case-cohort design [29], which, similar to the nested case-control design, is an efficient alternative to the full cohort design for the study of rare diseases. Intuitively, various partial-likelihood-based approaches that have been proposed for analyzing the case-cohort design under the standard Cox proportional hazards model should also be applicable for the induced hazard model that we have used in this article. Further work is needed for a formal theoretical development.

Acknowledgement

The authors wish to thank Kshama Aswath for helping resequence the control DNA population and B.J. Stone for helping improve the writing of the manuscript.

Appendix

For development of the asymptotic theory, we will assume the nested case-control design ($R_k^* = \tilde{R}_k$), but note that the proof for the cohort design follows by almost identical arguments.

A1: Derivation of Induced Hazard Function $\lambda\{t|G, \tilde{Z}(t)$

We derive the induced hazard of disease conditional on the genotype data G under the general situation when external time-dependent covariates $Z(t)$ are involved. We assume that D/G and $Z(t)$ are independent. The conditional hazard function can be obtained via $\lambda\{T|G, \tilde{Z}(T)\} = f\{t|G, \tilde{Z}(t)\}/\mathrm{p}\{T \geq t|G, \tilde{Z}(t)\}$. Note $\mathrm{p}\{T \geq t|G, \tilde{Z}(t)\} = \Sigma_{D \in D_G}\ \mathrm{p}\{T \geq t|D, \tilde{Z}(t)\}\mathrm{pr}_{\mathbf{f}}(D)/\mathrm{p}(G)$ and $f\{T|G, \tilde{Z}(t)\} = \Sigma_{D \in D_G}\ \lambda_0(t)e^{\beta^T\psi(D)+\gamma^TZ(t)}\ \mathrm{p}\{T \geq t|D, \tilde{Z}(t)\}\mathrm{pr}_{\mathbf{f}}(D)/\mathrm{p}(G)$. Thus, $\lambda\{t|G, \tilde{Z}(t)\} = \lambda_0(t)r_G\{t; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}$. This is the end of the proof.

A2: Derivation of $U_\beta(\mathbf{f}, \lambda)$ under the Null Hypothesis

We derive the score functions $U_\beta(\mathbf{f}, \gamma)$ under the null hypothesis $\beta = 0$. The log partial likelihood function is $\log L = \Sigma_{k:\Delta_k=1}\ [\log r_{G_k}\{t_k; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\} - \log \Sigma_{l \in \tilde{R}_k}\ r_{G_l}\{t_k; \mathbf{f}, \beta, \gamma, \lambda_0(\cdot)\}]$. Note that

$$\left.\frac{\partial r_G}{\partial \beta}\right|_{\beta=0}$$

$$= e^{\gamma^TZ(t)}\left.\frac{\Sigma_{D \in D_G}\ \psi(D)\mathrm{p}\{T \geq t|D, \tilde{Z}(t)\}\mathrm{pr}_{\mathbf{f}}(D)}{\Sigma_{D \in D_G}\ \mathrm{p}\{T \geq t|D, \tilde{Z}(t)\}\mathrm{pr}_{\mathbf{f}}(D)}\right|_{\beta=0}$$

$$= e^{\gamma^TZ(t)}\left.\frac{\Sigma_{D \in D_G}\ \psi(D)\mathrm{pr}_{\mathbf{f}}(D)}{\Sigma_{D \in D_G}\ \mathrm{pr}_{\mathbf{f}}(D)}\right|_{\beta=0} = e^{\gamma^TZ(t)}\,\mathrm{E}\{\psi(D)|G\}$$

and that $r_{G|_{\beta=0}} = e^{\gamma^TZ(t)}$. Then taking derivative of $\log L$ over β at $\beta = 0$ gives the desired score function for β. The score function for γ under the null hypothesis can simply be calculated.

A3: The Asymptotic Distribution of $U_\beta(\hat{\mathbf{f}}, \hat{\lambda})$

Denote $\mathrm{E}\{\psi(D)|G_i\}|_{\mathbf{f}=\mathbf{f}^*}$ by $o_{\mathbf{f}^*}(G_i)$. $I_{\beta\gamma}(\hat{\mathbf{f}}, \hat{\gamma})$ and $I_{\gamma\gamma}(\hat{\mathbf{f}}, \hat{\gamma})$ can be directly obtained from the derivatives of the partial likelihood function:

$$I_{\beta\gamma}(\hat{\mathbf{f}}, \hat{\gamma}) = \frac{1}{n}\sum_{k:\Delta_k=1}\left[\frac{\Sigma_{j \in \tilde{R}_k} o_{\hat{\mathbf{f}}}(G_j)Z_j^T e^{\hat{\gamma}^TZ_j(u)}}{\Sigma_{j \in \tilde{R}_k} e^{\hat{\gamma}^TZ_j(u)}} - \frac{\Sigma_{j \in \tilde{R}_k} o_{\hat{\mathbf{f}}}(G_j)\, e^{\hat{\gamma}^TZ_j(u)}\,\Sigma_{j \in \tilde{R}_k} Z_j^T e^{\hat{\gamma}^TZ_j(u)}}{\{\Sigma_{j \in \tilde{R}_k} e^{\hat{\gamma}^TZ_j(u)}\}^2}\right] \quad (12)$$

and

$$I_{\gamma\gamma}(\hat{\mathbf{f}}, \hat{\gamma}) = \frac{1}{n}\sum_{k:\Delta_k=1}\left[\frac{\Sigma_{j \in \tilde{R}_k} Z_j^{\otimes 2} e^{\hat{\gamma}^TZ_j(u)}}{\Sigma_{j \in \tilde{R}_k} e^{\hat{\gamma}^TZ_j(u)}} - \left\{\frac{\Sigma_{j \in \tilde{R}_k} Z_j\, e^{\hat{\gamma}^TZ_j(u)}}{\{\Sigma_{j \in \tilde{R}_k} e^{\hat{\gamma}^TZ_j(u)}\}}\right\}^{\otimes 2}\right], \quad (13)$$

where a vector with a superscript $\otimes 2$ denotes the vector multiplied by its transpose.

A3.1: Proof that $n^{-1}\,\partial U_\beta(\mathbf{f}^*, \gamma^*)/\partial \mathbf{f} \to 0$ for $(\mathbf{f}^*, \gamma^*)$ between $(\mathbf{f}, \gamma)$ and $(\hat{\mathbf{f}}, \hat{\gamma})$.

Let $N_i(t) = \Delta_i I(X_i \le t)$ be the counting process. Then $M_i(t) = N_i(t) - \int_0^t Y_i(u) e^{\gamma^T Z_i(u)} \lambda_0(u) du$ is a martingale. If $S_{\mathbf{f}^*}(G_i)$ denotes $\partial \mathrm{E}\{\psi(D)\,|G_i\}/\partial \mathbf{f}|_{\mathbf{f}=\mathbf{f}^*}$,

$$\mathrm{E}_{\mathbf{f}^*}(u) = \frac{\Sigma_{l \in R(u)} S_{\mathbf{f}^*}(G_l)\, e^{\gamma^T Z_l(u)}}{\Sigma_{l \in R(u)} e^{\gamma^T Z_l(u)}}$$

and

$$\tilde{\mathrm{E}}_{\mathbf{f}^*}(u) = \frac{\Sigma_{l \in \tilde{R}(u)} S_{\mathbf{f}^*}(G_l)\, e^{\gamma^T Z_l(u)}}{\Sigma_{l \in \tilde{R}(u)} e^{\gamma^T Z_l(u)}},$$

then

$$n^{-1} \frac{\partial U_\beta(\mathbf{f}^*, \gamma)}{\partial \mathbf{f}}$$

can be decomposed as

$$\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left[\left\{ S_{\mathbf{f}^*}(G_i) - \frac{\Sigma_{l \in \tilde{R}(u)} e^{\gamma^T Z_i(u)} S_{\mathbf{f}^*}(G_l)}{\Sigma_{l \in \tilde{R}(u)} e^{\gamma^T Z_i(u)}} \right\} dM_i(u) + \right.$$

$$\left. \{\mathrm{E}_{\mathbf{f}^*}(u) - \tilde{\mathrm{E}}_{\mathbf{f}^*}(u)\} Y_i(u)\, e^{\gamma^T Z_i(u)}\, \lambda_0(u) du \right],$$

where τ is the time of study termination. A slight modification of the results in Goldstein and Langholz [30] can be made to show that the first term is a martingale and the second term asymptotically converges to zero. Thus,

$$\frac{1}{n} \frac{\partial U_\beta(\mathbf{f}^*, \gamma^*)}{\partial \mathbf{f}} = \frac{1}{n} \frac{\partial U_\beta(\mathbf{f}^*, \gamma)}{\partial \mathbf{f}} + \left\{ \frac{1}{n} \frac{\partial U_\beta(\mathbf{f}^*, \gamma^*)}{\partial \mathbf{f}} - \frac{1}{n} \frac{\partial U_\beta(\mathbf{f}^*, \gamma)}{\partial \mathbf{f}} \right\} \to 0$$

A3.2: Derivation of Representation (10)

Now we derive the asymptotic distribution of the score functions for β under the null hypothesis $\beta = 0$. From the Taylor expansion for $U_\beta(\hat{\mathbf{f}}, \hat{\gamma})$ and using results in A3.1, $n^{-1/2}\, U_\beta(\hat{\mathbf{f}}, \hat{\gamma}) \approx n^{-1/2}\, U_\beta(\mathrm{f}, \gamma) - i_{\beta\gamma} \sqrt{n}(\hat{\gamma} - \gamma)$, where $(\mathbf{f}^*, \gamma^*)$ is between (f, γ) and $(\hat{\mathbf{f}}, \hat{\gamma})$ and thus $(\mathbf{f}^*, \gamma^*) \to (\mathbf{f}, \gamma)$. Using the Taylor expansion for $U_\gamma(\hat{\gamma})$ yields $\sqrt{n}\,(\hat{\gamma} - \gamma) \approx i_{\gamma\gamma}^{-1}\, n^{-1/2}\, U_\gamma(\gamma)$. Putting the above equations together leads to

$$\frac{1}{\sqrt{n}} U_\beta(\hat{\mathbf{f}}, \hat{\gamma}) \approx \frac{1}{\sqrt{n}} U_\beta(\mathbf{f}, \gamma) - i_{\beta\gamma} i_{\gamma\gamma}^{-1} \frac{1}{\sqrt{n}} U_\gamma(\gamma)$$

$$\sim N(0, i_{\beta\beta} - i_{\beta\gamma} i_{\gamma\gamma}^{-1} i_{\beta\gamma}^T)$$

A3.3: Derivation of the Estimator for $i_{\beta\beta}$ under the Null Hypothesis

The derivation follows Goldstein and Langholz [30] and Xiang and Langholz [31].

$$\frac{1}{\sqrt{n}} U_\beta(\mathbf{f}^*, \gamma) \approx$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \left\{ o_{\mathbf{f}^*}(G_i) - \frac{\Sigma_{l \in \tilde{R}_i(u)} e^{\gamma^T Z_i(u)}\, o_{\mathbf{f}^*}(G_l)}{\Sigma_{l \in \tilde{R}_i(u)} e^{\gamma^T Z_i(u)}} \right\} dM_i(u).$$

Let

$$H = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \left\{ o_{\mathbf{f}^*}(G_i) - \frac{\Sigma_{l \in \tilde{R}_i(u)} e^{\gamma^T Z_i(u)}\, o_{\mathbf{f}^*}(G_l)}{\Sigma_{l \in \tilde{R}_i(u)} e^{\gamma^T Z_i(u)}} \right\} dM_i(u),$$

which is square-integrable martingale with terms in the sum uncorrelated. Thus $H \sim N(0, <H, H>)$. Following Langholz [32], $i_{\beta\beta}$, which is equal to $<H, H>$, can be simply estimated based on an information matrix of the form

$$I_{\beta\beta}(\hat{\gamma}, \hat{\mathbf{f}}) = \frac{1}{n} \sum_{k:\Delta_k = 1} \left\{ \frac{\Sigma_{j \in \tilde{R}_k} o_{\hat{\mathbf{f}}}^{\otimes 2}(G_j)\, e^{\gamma^T Z_i(u)}}{\Sigma_{j \in \tilde{R}_k} e^{\gamma^T Z_i(u)}} \right.$$

$$\left. - \left(\frac{\Sigma_{j \in \tilde{R}_k} o_{\hat{\mathbf{f}}}^{\otimes 2}(G_j)\, e^{\gamma^T Z_i(u)}}{\Sigma_{j \in \tilde{R}_k} e^{\gamma^T Z_i(u)}} \right)^{\otimes 2} \right\}. \qquad (14)$$

References

1 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002; 70:425–434.

2 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003;33:228–237.

3 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:1516–1517.

4 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003; 73:1316–1329.

5 Fallin D, Schork N: Accuracy of haplotype frequency estimation for biallelic loci via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 2000;67:947–959.

6 Wallenstein S, Hodge S, Weston A: A logistic regression model for analyzing extended haplotype data. Genet Epidemiol 1998;15:173–181.

7 Stram D, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC: Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 2003;55: 179–190.

8 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002;53: 79–91.

9 Thomas DC: Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining, by FDK Liddell, JC McDonald, and DC Thomas. J R Stat Soc, series A, 1977;140: 469–491.

10 Hu FB, Doria A, Li T, Meigs JB, Liu S, Memisoglu A, Hunter D, Manson JE: Genetic variation at the adiponectin locus and risk of type 2 diabetes in women. Diabetes 2004;53:209–213.

11 Setiawan VW, Hankinson SE, Colditz GA, Hunter DJ, De Vivo I: HSD17B1 gene polymorphisms and risk of endometrial and breast cancer. Cancer Epidemiol, Biomarkers Prevention 2004;13(2):213–219.

12 Mohlig M, Boeing H, Spranger J, Osterhoff M, Kroke A, Fisher E, Bergmann MM, Ristow M, Hoffmann K, Pfeiffer AF: Body mass index and C-174G interleukin-6 promoter polymorphism interact in predicting type 2 diabetes. J Clin Endocrinol Metab 2004;89:1885–1890.

13 Nieters A, Linseisen J, Becker N: Association of polymorphisms in Th1, Th2 cytokine genes with hayfever and atopy in a subsample of EPIC-Heidelberg. Clin Exp Allergy 2004;34: 346–353.

14 Haiman CA, Stram DO, Pike MC, Kolonel LN, Burtt NP, Altshuler D, Hirschhorn J, Henderson BE: A comprehensive haplotype analysis of CYP19 and breast cancer risk: The Multiethnic Cohort. Hum Mol Genet 2003;12:2679–2692.
15 Paltoo D, Woodson K, Taylor P, Albanes D, Virtamo J, Tangrea J: Pro12Ala polymorphism in the peroxisome proliferator-activated receptor-gamma (PPAR-gamma) gene and risk of prostate cancer among men in a large cancer prevention study. Cancer Lett 2003;191:67–74.
16 Woodson K, Ratnasinghe D, Bhat NK, Stewart C, Tangrea JA, Hartman TJ, Stolzenberg-Solomon R, Virtamo J, Taylor PR, Albanes D: Prevalence of disease-related DNA polymorphisms among participants in a large cancer prevention trial. Eur J Cancer Prevention 1999;8:441–447.
17 Giovannucci E, Chen J, Smith-Warner SA, Rimm EB, Fuchs CS, Palomeque C, Willett WC, Hunter DJ: Methylenetetrahydrofolate reductase, alcohol dehydrogenase, diet, and risk of colorectal adenomas. Cancer Epidemiol Biomarkers Prevention 2003;12:970–979.
18 Hsu L: Genetic association tests with age at onset. Genet Epidemiol 2003;24:118–127.
19 Zhao LP, Li S, Khalid N: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 2003;72:1231–1250.
20 Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork N: Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 2001;11:143–151.
21 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921–927.
22 Long J, Williams R, Urbanek M: An EM algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995;56:799–810.
23 Niu T, Qin Z, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. Am J Hum Genet 2002;70:157–169.
24 Fleming TR, Harrington DP: Counting processes and survival analysis. New York, John Wiley & Sons, 1991.
25 Hosmer DW, Lemeshow S: Applied survival analysis: regression modeling of time to event data. New York, John Wiley & Sons, 1999.
26 Borgan O, Goldstein L, Langholz B: Methods for the analysis of sampled cohort data in the Cox proportional hazards model. Ann Stat 1995;23:1749–1778.
27 Lin DY: Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol 2004;26(4):255–264.
28 Sasieni P: Maximum weighted partial likelihood estimators for the Cox model. J Am Stat Assoc 1993;88:144–152.
29 Prentice, RL: A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.
30 Goldstein L, Langholz B: Asymptotic theory for nested case-control sampling in the Cox regression model. Ann Stat 1992;20:1903–1928.
31 Xiang AH, Langholz B: Robust variance estimation for rate ratio parameter estimates from individually matched case-control data. Biometrika 2003;90:741–746.
32 Langholz B: Robust variance estimation for rate ratio parameter estimates from individually matched case-control data: Supplemental material. Technical Report, Department of Preventive Medicine, University of Southern California, 2003.
33 Cordell HJ, Clayton DG: A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. Am J Hum Genet 2002;70:124–141.